Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

IR-SegDiff: Conditional Diffusion for Subsurface Defect Segmentation and Uncertainty Estimation in Infrared Thermography

Jin Nan
Beihang University
Chaoqi Huang
Beihang University
Shunmin Wu
Beihang University
Qi Zhao
Beihang University
Zhijun Xu
Hangzhou Hikrobot Co., Ltd.
Min Li
Beihang University
Yuhang Li
Beihang University
Yunpeng Ma
myp@buaa.edu.cn
Beihang University

Research Article

Keywords: Infrared thermography, subsurface defect segmentation, conditional diffusion model, multi-scale feature fusion, uncertainty estimation

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11581386/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

IR-SegDiff: Conditional Diffusion for Subsurface Defect Segmentation and Uncertainty Estimation in Infrared Thermography

Jin Nan[1] , Chaoqi Huang[1] , Shunmin Wu[1] , Qi Zhao[1] , Zhijun Xu[2], Min Li[1] , Yuhang Li[1,4,5*], Yunpeng Ma[3*]

1. Institute of Solid Mechanics, Beihang University, Beijing, 100191, China.
2. Hangzhou Hikrobot Co., Ltd., Hangzhou, 310051, China.
3. School of Aeronautic Science and Engineering, Beihang University, Beijing, 100191, China.
4. Tianmushan Laboratory, Xixi Octagon City, Yuhang District, Hangzhou 310023, China.
5. Aircraft and Propulsion Laboratory, Ningbo Institute of Technology, Beihang University, Ningbo, 315800, China.

nanjin@buaa.edu.cn (Jin Nan); zy2405201@buaa.edu.cn (Chaoqi Huang); wsm420881@buaa.edu.cn (Shunmin Wu); zqibj@163.com (Qi Zhao); xuzhijun@hikrobotics.com (Zhijun Xu) ;limin@buaa.edu.cn (Min Li) ; liyuhang@buaa.edu.cn (Yuhang Li).
*Correspondence: myp@buaa.edu.cn (Yunpeng Ma).

Abstract

Automatic detection of subsurface defects in aerospace structures is critical for structural safety assessment. In infrared thermography, the low thermal contrast of subsurface defects and measurement disturbances complicate defect segmentation and boundary delineation, while a single segmentation result provides limited information about potential prediction errors. To address these challenges, this study proposes IR-SegDiff, a conditional diffusion-based segmentation model for subsurface defect detection in structural materials commonly used in aerospace applications. The model integrates a conditional encoder, a conditional cross-attention fusion module (CCAFM), and a multi-scale feature fusion module (MSFFM) to guide defect reconstruction. During inference, DPM-Solver accelerates the sampling process, while repeated sampling enables ensemble prediction and pixel-wise uncertainty characterization. On a self-constructed dataset comprising carbon fiber-reinforced polymer (CFRP), glass fiber-reinforced polymer (GFRP), and resin specimens, the five-sample ensemble achieves an $F1$-Score of 76.72% and an intersection over union (IoU) of 65.43%, outperforming seven baseline segmentation networks. The 25-step DPM-Solver achieves a 35.57-fold speedup compared with conventional step-by-step denoising. Controlled noise experiments demonstrate that ensemble sampling improves segmentation performance under noisy conditions, while sampling-based uncertainty provides complementary information for false-negative localization. By integrating condition-guided segmentation, efficient inference, and uncertainty characterization, the proposed method provides an approach for defect identification and assisted interpretation in infrared thermography.

Keywords: Infrared thermography; subsurface defect segmentation; conditional

diffusion model; multi-scale feature fusion; uncertainty estimation.

1. Introduction

With the increasing demand for lightweight and high-strength aerospace structures, various materials, including composites, have been widely employed in critical load-bearing components [1]. During long-term service, these materials may develop internal defects, such as cracks, delaminations, and voids, posing potential threats to structural integrity and safety [2]. To address defect detection requirements across different material systems, various nondestructive testing (NDT) techniques have been employed, including eddy current testing [3], ultrasonic testing [4],[5], radiographic testing [6], and magnetic particle testing [7]. Infrared thermography (IRT), as a non-contact and rapidly deployable NDT technique with relatively broad material applicability, provides an effective approach for detecting internal defects in aerospace structures [8]. However, IRT is susceptible to noise introduced by environmental disturbances and variations in testing conditions, while defect identification accuracy remains dependent on operator experience. Therefore, developing reliable and effective automated visual inspection methods is essential for improving defect detection performance.

Conventional signal processing methods enhance defect features and suppress noise in infrared thermography [9],[10]. Fourier transform [11] and wavelet transform [12] exploit phase information to mitigate nonuniform heating and surface emissivity variations. Principal component analysis (PCA) [13] and thermographic signal reconstruction (TSR) [14] highlight defects through thermal response feature extraction or reconstruction, while dual-orthogonal demodulation (DOD) jointly extracts amplitude and phase information [15]. Although these methods improve defect visibility, their reliance on predefined signal processing models limits the representation of nonlinear, multiscale defect features under complex noise and varying operating conditions.

Deep learning has driven the development of infrared thermography-based defect detection from handcrafted feature analysis toward data-driven modeling [16]. Yao et al. [17] combined future-frame prediction with polarization information to extract defect-related features. Cheng et al. [18] and Matarrese et al. [19] employed multitask learning and temporal convolutional networks, respectively, for defect identification and depth characterization. To improve input quality, Morelli et al. [20] combined sequence preprocessing with a residual network for segmentation under varying heating conditions, while Lim et al. [21] and Li et al. [22] used physics-informed neural networks and denoising autoencoders for defect feature enhancement. For data augmentation, Fang et al. [23] integrated generative adversarial networks with partial least squares thermography for image enhancement and noise suppression. These studies advance automated infrared defect identification through feature learning and data enhancement.

Building on advances in infrared thermographic data augmentation and enhancement, generative models have provided new approaches to defect detection

through sample synthesis, image enhancement, and defect reconstruction [24]. In infrared thermography, Cheng et al. proposed Dual-IRT-GAN[25] and IRT-GAN [26], incorporating super-resolution reconstruction and multi-head feature fusion for infrared defect enhancement and segmentation, respectively. Denoising diffusion probabilistic models (DDPMs) provide an alternative for defect reconstruction and sample generation through progressive denoising [28],[29]. Zhou et al. [27] proposed DiffDD, combining multiscale PVTv2 features with diffusion-based reconstruction for surface defect segmentation. Xia et al. [30] introduced attention mechanisms in DASeg-Net to capture local details and global dependencies. Fang et al. [31] integrated conditional diffusion with human feedback and reinforcement learning for controllable defect generation. These studies demonstrate the potential of diffusion models for accurate defect detection, controllable generation, and feature enhancement in industrial inspection.

Given the advantages demonstrated by diffusion models in industrial defect detection, their application to infrared thermography has also attracted increasing research interest. Wei et al. [32] proposed a DDPM-based framework that outperformed CVAE and CGAN models on composite specimens. These frameworks primarily incorporate handcrafted features, such as ICA, PCA, and PLSR, as conditional constraints through a conditional encoder. Such relatively shallow conditional feature integration limits multi-scale feature interactions, making it difficult to simultaneously preserve defect structures, delineate boundary details, and distinguish defects from the background. Furthermore, conventional DDPMs require substantial computational resources during inference, limiting their efficiency in practical inspection applications. Existing studies have also largely treated diffusion models as deterministic segmentation tools, focusing primarily on final segmentation accuracy without fully exploiting prediction variability to identify potential false negatives.

To address these limitations, this study proposes IR-SegDiff, a conditional diffusion-based segmentation model for subsurface defect detection in infrared thermography. The model employs a conditional encoder to extract multi-scale features from infrared images and a CCAFM to integrate conditional information with noisy segmentation features, guiding the progressive reconstruction of defect regions. A MSFFM further integrates structural and contextual information across different feature levels, enhancing defect representation and discrimination from the background. During inference, DPM-Solver is adopted to reduce sampling costs, while repeated sampling generates ensemble segmentation predictions and pixel-wise uncertainty estimates, providing additional information for defect localization and the review of potential false-negative regions. Using a custom-built active infrared thermography system, a dataset covering different materials, defect configurations, and acquisition conditions is constructed. Comprehensive experiments are conducted, including comparisons with baseline models, module ablation studies, sampling efficiency evaluation, and controlled noise experiments. The main contributions of this study are summarized as follows:

1) A conditional diffusion-based segmentation model, IR-SegDiff, is proposed to integrate infrared image-guided defect mask generation with repeated

sampling ensemble, enabling pixel-level localization of subsurface defects. Its segmentation performance is validated on a test set partitioned by physical specimens.

2) A multi-scale conditional guidance and feature fusion mechanism is developed to incorporate infrared defect information into the diffusion denoising process through the conditional encoder, CCAFM, and MSFFM, improving defect feature representation and enhancing the suppression of false positives in background regions.
3) DPM-Solver accelerated sampling is combined with predictive uncertainty analysis to reduce inference costs while generating ensemble predictions and pixel-wise prediction variability information. Controlled noise experiments and evaluation of false-negative capture rates at a fixed pixel review budget assess its utility for defect segmentation and prioritizing regions for review under noisy conditions.

2. Methodology

2.1. Principle and system of infrared thermographic inspection

Infrared thermographic testing identifies and localizes internal defects by capturing infrared radiation emitted from the surface of the inspected structure and analyzing the temporal characteristics of its thermal response. This study employs two active modulated thermal excitation methods: single-frequency sinusoidal excitation and linear frequency modulation (LFM). Single-frequency sinusoidal excitation applies a sinusoidally modulated heat source at a fixed frequency, followed by phase and amplitude demodulation of the thermal response to improve the signal-to-noise ratio of deeply embedded defects [33]. In contrast, LFM employs a frequency-swept excitation with continuously varying frequency, covering multiple frequencies within a single experiment to characterize defects at different depths [34]. Both methods are based on thermal wave propagation governed by the heat diffusion equation. By adjusting the modulation frequency, the characteristic thermal diffusion length can be varied, facilitating the detection of defects with different sizes and burial depths.

To ensure stable and controllable active thermal excitation and high-precision data acquisition under modulated excitation conditions, an active infrared thermography inspection system was developed. As shown in Fig. 1(a), the system consists of a laser heat source, a modulation control unit, an infrared camera, and a data acquisition and processing unit. A control program coordinates the thermal excitation and image acquisition processes. The laser beam illuminates the specimen surface through a diffuser, while the infrared camera records the temporal temperature evolution during heating. The acquired thermal image sequences are subsequently used for data preprocessing and defect segmentation.

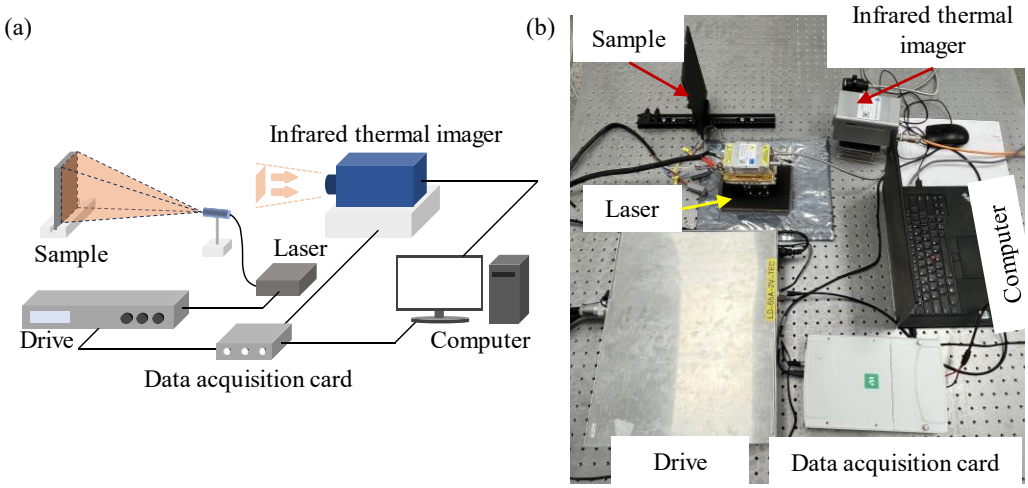


Fig. 1. Infrared thermography detection system. (a) Infrared heating and signal acquisition process; (b) Experimental system.

2.2. IR-SegDiff model

To address the limited detection accuracy and stability of infrared thermography under varying testing conditions, this study proposes IR-SegDiff, a diffusion-based model guided by multi-feature fusion for subsurface defect detection, as illustrated in Fig. 2. Built upon a conditional diffusion probabilistic framework, the model extracts multi-scale semantic features from input infrared images to guide the diffusion denoising process, enabling pixel-wise defect segmentation.

2.2.1. Overall architecture

The proposed diffusion-based segmentation model, IR-SegDiff, consists of a conditional encoder and a denoising network, incorporating conditional cross-attention fusion modules (CCAFM)[35] and a multi-scale feature fusion module (MSFFM). At the diffusion timestep t, the model takes the preprocessed infrared image b and the same-sized noisy segmentation map x_t as inputs, and predicts the noise component $\hat{\varepsilon}_t$.

The denoising backbone adopts a U-Net-based noise predictor to learn the conditional distribution of the reverse diffusion process. To capture and incorporate detailed information from the input infrared image, a conditional encoder extracts hierarchical features from b at multiple spatial scales. These features are integrated with the corresponding encoder features of the U-Net backbone through CCAFM using cross-attention. During decoding, multi-scale features from the backbone are aggregated through MSFFM to integrate structural and contextual information. The resulting fused features are used to predict the noise component at the current timestep, enabling a single update in the reverse diffusion process.

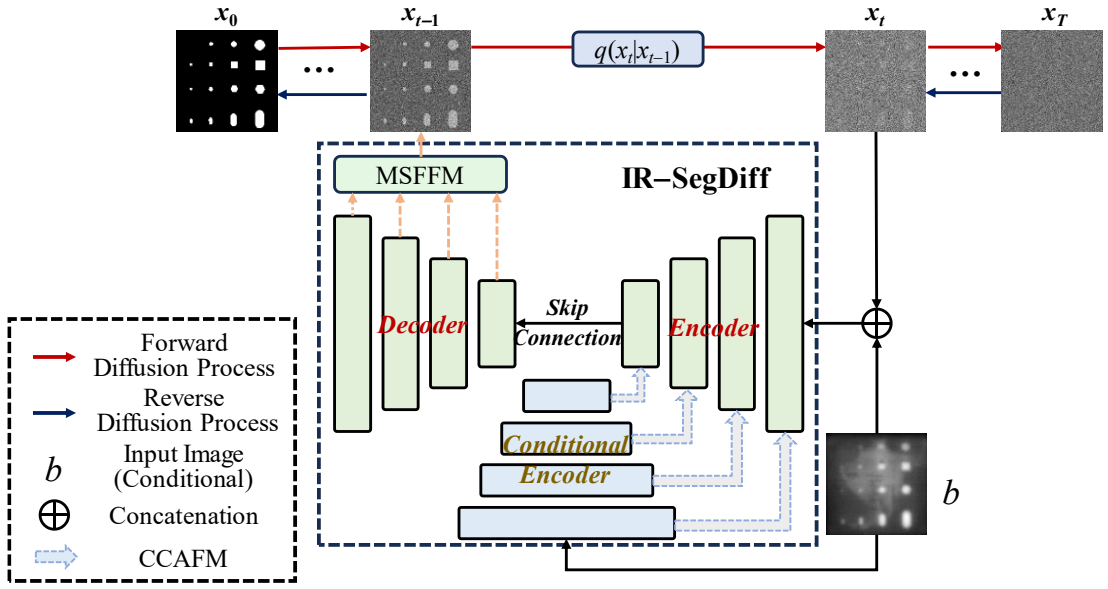


Fig. 2. The overview of the proposed IR-SegDiff.

2.2.2. Denoising module

The IR-SegDiff diffusion model consists of two stages: a forward diffusion process and a reverse generative process. In the forward process, Gaussian noise is progressively injected into the ground-truth segmentation map x_0 to obtain a noisy sample x_t, which can be formulated as:

$$q\left(x_t|x_{t-1}\right)=N\left(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t I\right),\quad \forall t\in\{1,\cdots,T\}, \tag{1}$$

where t denotes the timestep in the Markov chain, $\beta_t\in[0,1]$ is the diffusion variance, and I is the identity matrix. Combining $\alpha_t=1-\beta_t$, $\bar{\alpha}_t=\prod_{s=1}^{t}\alpha_s$, and $\varepsilon\sim N(0,I)$, x_t at an arbitrary timestep can be sampled directly from x_0:

$$x_t=\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\varepsilon. \tag{2}$$

During training, parameter learning is conducted within the DDPM framework [28]. The model predicts the noise component ε at time step t to minimize the mean squared error (MSE) loss function:

$$L_{DDPM}=E_{x_0,t,\varepsilon}\left[\left\|\varepsilon-\varepsilon_\theta\left(x_t,t,b\right)\right\|^2\right], \tag{3}$$

ε_θ denotes the parameterized noise prediction network. By learning the denoising mappings across noise levels, the model progressively reconstructs the target segmentation map during the reverse diffusion process.

During inference, to improve sampling efficiency and reduce computational cost, DPM-Solver is adopted to accelerate sampling of the reverse diffusion process[36]. This method transforms the stochastic differential equation (SDE) that describes the reverse process of the continuous time diffusion model into an equivalent probability flow ordinary differential equation (ODE):

$$\frac{dx_t}{dt} = f(t)x_t + \frac{g^2(t)}{2\sigma_t}\varepsilon_\theta(x_t, t), x_T \square q_T(x_T), \tag{4}$$

where $\varepsilon_\theta(x_t,t)$ denotes the noise predicted by the network and x_t represents the noisy segmentation map at timestep t. The resulting ODE describes a deterministic evolution trajectory of samples in continuous time. For efficient numerical solving, following the standard DPM-Solver formulation, the time variable is reparametrized by the log signal-to-noise ratio λ, and the ODE can be expressed in an integral form with respect to λ:

$$x_t = \frac{\alpha_t}{\alpha_s}x_s - \alpha_t\int_{\lambda_s}^{\lambda_t} e^{-\lambda}\hat{\varepsilon}_\theta(\hat{x}_\lambda, \lambda)d\lambda. \tag{5}$$

By applying a (k-1)-order Taylor expansion of the noise prediction function around $\lambda_{t_{i-1}}$, the integral term is approximated with higher-order accuracy, yielding the k-th order update rule:

$$x_{t_{i-1}\to t_i} = \frac{\alpha_{t_i}}{\alpha_{t_{i-1}}}x_{t_{i-1}} - \alpha_{t_i}\sum_{n=0}^{k-1}\varepsilon_\theta^{(n)}(x_{\lambda_{t_{i-1}}}, \lambda_{t_{i-1}})\int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} e^{-\lambda}\frac{(\lambda-\lambda_{t_{i-1}})^n}{n!}d\lambda + \mathrm{O}(h_i^{k+1}) \tag{6}$$

where $\varepsilon_\theta^{(n)}$ denotes the n-th derivative with respect to λ, $h=\lambda_{t_i}-\lambda_{t_{i-1}}$ is the step size, and $\mathrm{O}(h_i^{k+1})$ is the higher-order truncation error term. By constructing an explicit integral approximation using multi-order information, the k-th order DPM-Solver maintains numerical stability and sampling accuracy under relatively large step sizes.

2.2.3. Conditional cross-attention fusion module (CCAFM)

To enhance the responsiveness of the diffusion backbone to conditional information, a CCAFM is designed to provide multi-scale conditional guidance during encoding. As shown in Fig. 3, the module takes the backbone U-Net encoder feature $\varphi_i(x_t) \in R^{C_i\times H_i\times W_i}$ and the conditional feature $\psi_i(b) \in R^{C_i\times H_i\times W_i}$ extracted by the conditional encoder as inputs, where i=1，2，3，4 denotes the network hierarchy. As illustrated in Fig. 3(b), the two feature streams are first processed by cross-attention to capture their correlations. Specifically, linear projections generate the Query, Key, and Value vectors: $Q=W_q\varphi_i(x_t)$, $K=W_k\psi_i(b)$, $V=W_v\psi_i(b)$，where W_q，W_k，W_v are learnable projection matrices. The cross-attention output is then computed as:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{7}$$

where d is the feature dimension scaling factor. Through this mechanism, the backbone encoder features can adaptively select conditional cues that are most relevant to the current structure, providing semantic guidance for diffusion denoising. To further enhance the representational capacity of the fused features, an attention and convolution fusion module (ACFM) is employed[37]. ACFM operates on the cross-attention output and comprises a global branch and a local branch, which model long-range semantic

dependencies and local spatial details, respectively, as illustrated in Fig. 3(c). In the global branch, a self-attention mechanism is introduced to strengthen long-range interactions across the spatial domain. Unlike the aforementioned cross-attention, this self-attention operates solely on the fused features, modeling global relationships within the feature itself.

In the local branch, a 1×1 convolution is first applied for channel adjustment, followed by a channel shuffle operation that divides the features into several groups along the channel dimension. Depthwise separable convolution is then performed within each group to facilitate cross-channel information exchange and reduce parameter redundancy. Subsequently, the 3×3 convolution is employed to further extract local spatial features. Finally, the outputs of the global and local branches are fused to obtain an enhanced feature representation, which is then passed to subsequent network layers.

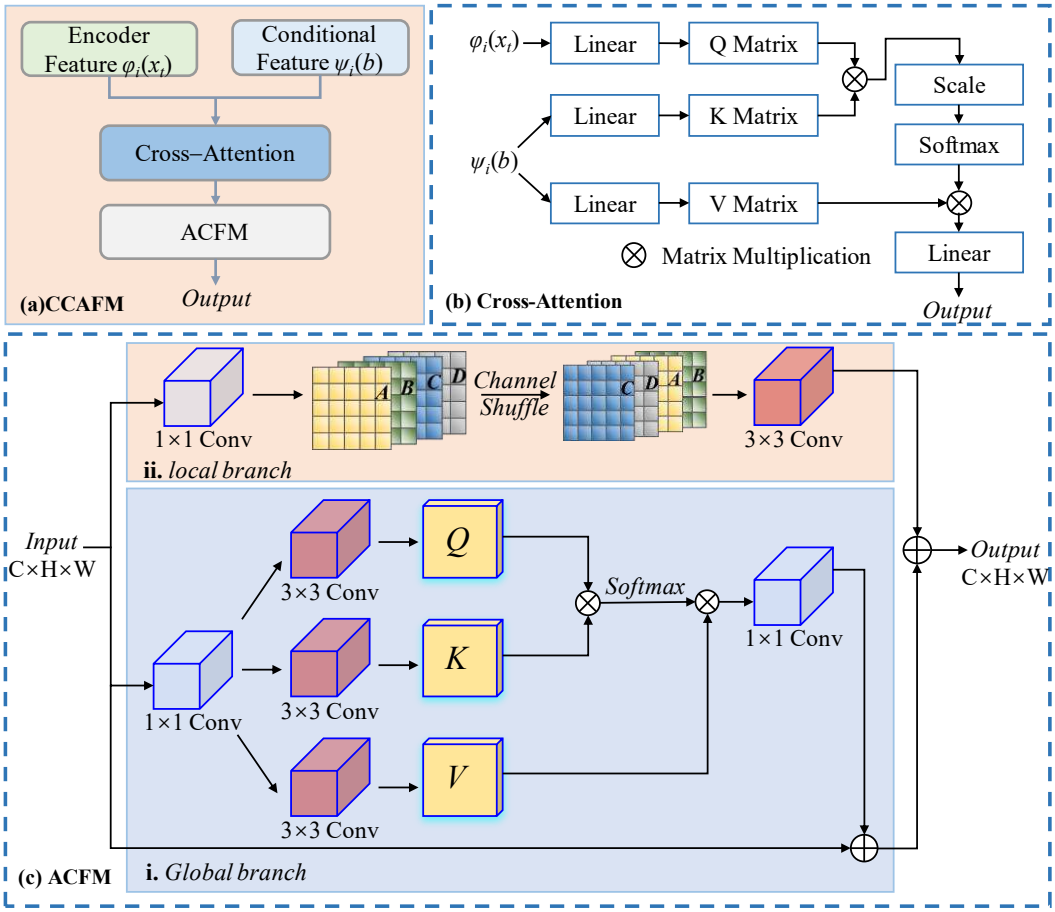


Fig. 3. The architecture of CCAFM. (a) CCAFM; (b) Cross-attention; (c) ACFM.

2.2.4. Multi-scale feature fusion module (MSFFM)

Multi-scale segmentation features contain rich structural and texture information. To fully exploit semantic information and fine-grained structural details across different levels of the decoding process, this study introduces a MSFFM. The module integrates multi-scale features extracted from different layers of the U-Net backbone decoder, enhancing the model's feature extraction and aggregation capabilities.

The architecture of the proposed MSFFM is illustrated in Fig. 4. As the backbone adopts a U-Net architecture, the spatial resolution of the feature maps is doubled at each

upsampling stage during decoding, progressively restoring the original resolution of 224 × 224. Since features at different scales vary in channel dimensions and spatial resolutions, 1 × 1 convolutions are first applied to map their channel dimensions to a uniform size before fusion. The resulting feature maps are then concatenated along the channel dimension, followed by a 3 × 3 convolution to integrate cross-scale information. Batch normalization and ReLU activation are subsequently employed to enhance nonlinear feature representation. Finally, average pooling is applied to impose a smoothing constraint, producing a denoised segmentation output with the same spatial dimensions as the input. By aggregating multi-scale features from the decoder, MSFFM preserves both high-level semantic information and low-level structural details, thereby improving the model's ability to identify infrared defects of varying sizes and delineate complex defect boundaries.

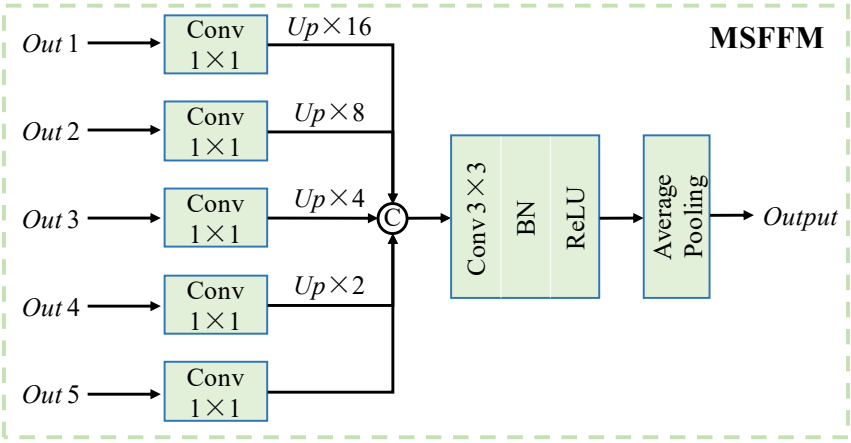


Fig. 4. The architecture of MSFFM.

3. Experimental campaign

3.1. Dataset

To evaluate the detection performance of the proposed method across different materials and defect configurations, an active infrared thermography inspection system was developed to experimentally investigate structural materials commonly used in aerospace applications. As shown in Fig. 1(b), the system employs a JOLD−45 laser operating at 808 nm as the thermal excitation source and a FLIR A615 infrared camera to record surface thermal responses at a spatial resolution of 640×480 pixels and a frame rate of 25 Hz. A control program developed using the device software development kit (SDK) synchronizes the start and stop of thermal excitation and image acquisition. The LFM excitation covers a frequency range of 0.01–0.1 Hz, while the single-frequency sinusoidal excitation operates at 0.1 Hz. Each acquisition lasts 100 s.

The dataset comprises eight specimens with artificial defects, including five CFRP specimens (C1–C5), two GFRP specimens (G1–G2), and one resin specimen (R1).Fig. 5 presents three-view geometrical illustrations of the eight specimens, while Table 1 summarizes their materials, in-plane defect geometries, burial depths, and configurations. Specimens C1, C2, and R1 contain circular flat-bottom holes with different diameters and burial depths, whereas C3 contains irregular flat-bottom defects. Specimens C4 and C5 contain stepped-depth defects and defects with irregular in-plane

contours and depth distributions, respectively. Specimens G1 and G2 feature flat-bottom holes and fully resin-filled defects, respectively. These specimens are used to evaluate segmentation performance under different material backgrounds and defect geometries.

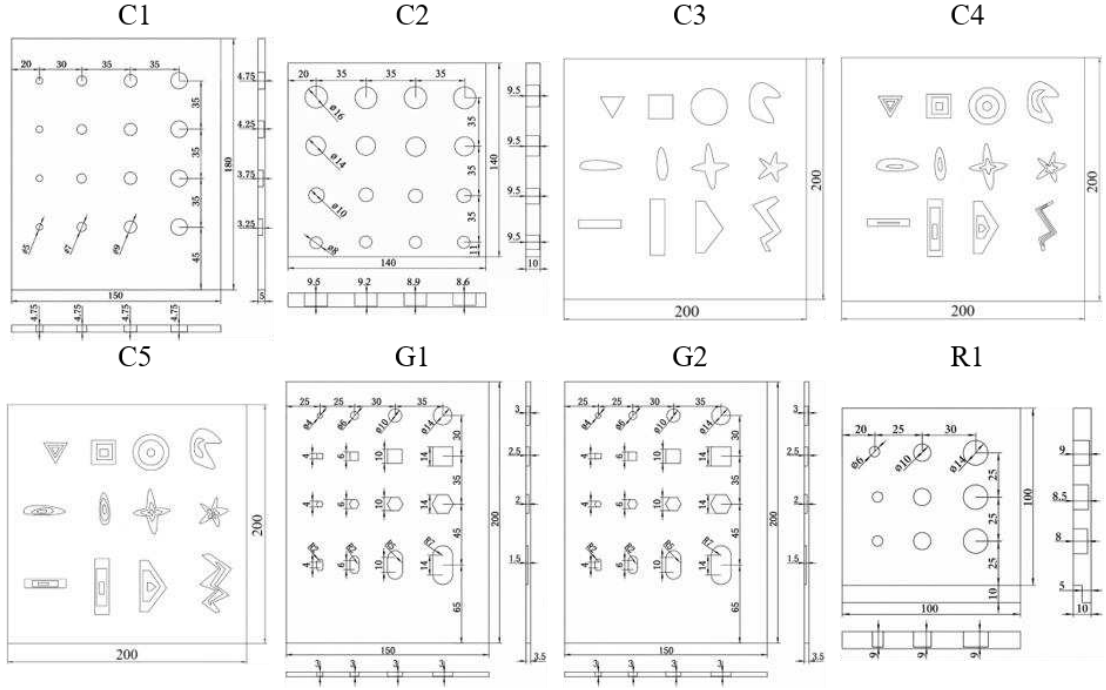


Fig. 5. Geometries and defect layouts of the eight specimens.

Table 1
Materials, defect geometries, and configurations of the eight specimens.

Specimen	Host material	In-plane defect geometry size (mm)	Defect depth below the inspected surface (mm)	Defect configuration
C1	CFRP	Circular, 5, 7, 9, 10	0.25, 0.75, 1.25, 1.75	Flat-bottom holes
C2	CFRP	Circular, 8, 10, 14, 16	0.5, 0.8, 1.1, 1.4	Flat-bottom holes
C3	CFRP	Irregular	2.0	Irregular flat-bottom defects
C4	CFRP	Irregular	Stepped profile, up to 2.0	Stepped-depth defects
C5	CFRP	Irregular	Irregular profile	Irregular-shape and irregular-depth defects
G1	GFRP	Circular, 4, 6, 10, 14	0.5, 1.0, 1.5, 2.0	Flat-bottom holes
G2	GFRP	Circular, 4, 6, 10, 14	0.5, 1.0, 1.5, 2.0	Fully resin-filled defects
R1	Resin	Circular, 6, 8, 10	0.5, 1.0, 1.5	Flat-bottom holes

Each specimen was tested under three conditions for each of the two excitation modes, yielding six thermal image sequences per specimen. The three conditions included no additional disturbance, mild external disturbances, and specimen repositioning. External disturbances were introduced using airflow and an external

infrared heat source. For the repositioning condition, the specimen position was adjusted before acquisition and kept fixed throughout the acquisition period. A total of 48 infrared thermal image sequences were collected from the eight specimens. For each sequence, the temporal temperature response at each pixel was fitted using a second-order polynomial, and the fitted trend was subtracted to suppress the slowly varying background temperature rise. The 100s acquisition period was then divided into four consecutive 25s intervals, from each of which four frames were randomly selected, yielding 16 frames per sequence to cover different heating stages. Consequently, each specimen contributed 96 images, and the complete dataset comprised 768 image label pairs.

The segmentation labels were generated from the planar projections of defects in the three-dimensional specimen models and mapped to the corresponding infrared images after proportional scaling and spatial registration. The model input consisted of single-frame grayscale images subjected to temporal detrending. Representative samples are shown in Fig. 6. The dataset was partitioned by physical specimen: specimens C1, C2, C3, G1, G2, and R1 were used for training, comprising 36 sequences and 576 image-label pairs, while specimens C4 and C5 were reserved for testing, comprising 12 sequences and 192 image label pairs. All sequences, extracted frames, and derived cropped regions from the same specimen were assigned to the same subset to prevent data leakage between the training and test sets.

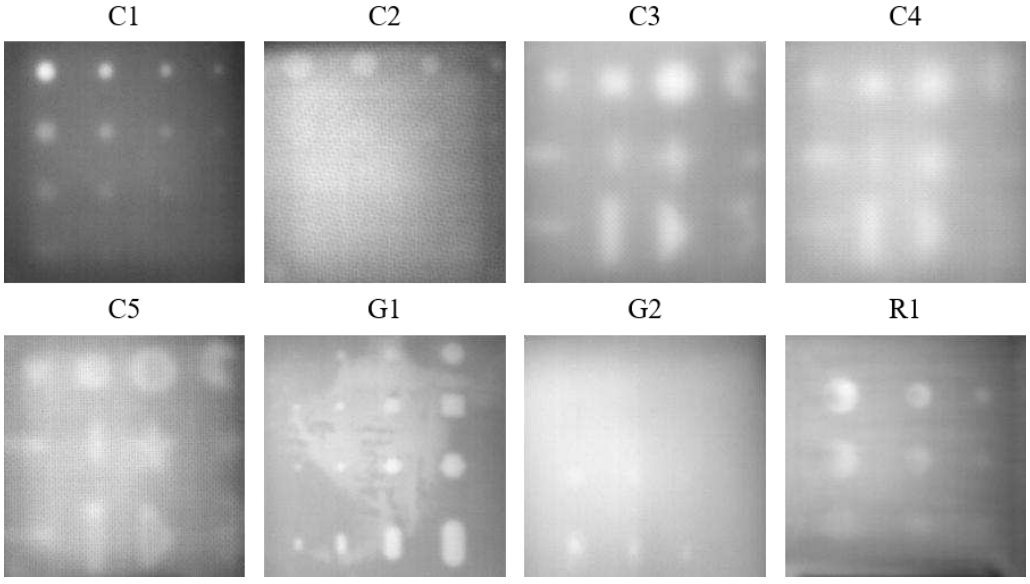


Fig. 6. Examples of the defect dataset comprising eight experimental groups.

3.2. Evaluation metrics

In this study, defect detection is formulated as a binary classification task, where the detection results are represented as binary masks. To comprehensively evaluate the proposed method, commonly used metrics in prior studies are adopted, including intersection over union (IoU), precision, Recall, and the *F*1-Score [17],[27],[38]. Specifically, IoU measures the overlap between the predicted segmentation and the ground truth, defined as the ratio of the intersection area to the union area of the two

regions, thereby quantifying their spatial agreement. Precision represents the proportion of true positives among all samples predicted as positive, reflecting the reliability of the predictions. Recall represents the proportion of actual positive samples correctly identified by the model, indicating its sensitivity to defects. The $F1$-Score is the harmonic mean of Precision and Recall, providing a balanced evaluation of segmentation performance by accounting for both false positives and false negatives. These metrics are calculated as follows:

$$F1\text{-Score}=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}, \tag{8}$$

$$\text{IoU}=\frac{\text{TP}}{\text{TP+FP+FN}}, \tag{9}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}}, \tag{10}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}}, \tag{11}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

To further assess the overall discriminative performance of the model across different thresholds, average precision (AP) is also reported [39]. Furthermore, given the sensitivity of infrared defect detection to false alarms, the $F_{0.5}$-Score, specificity, and false positive rate (FPR) are employed to evaluate the model's ability to control false positive predictions:

$$F_{0.5}\text{-Score}=\frac{1.25\times \text{Precision}\times \text{Recall}}{0.25\times \text{Precision}+\text{Recall}}, \tag{12}$$

$$\text{Specificity}=\frac{\text{TN}}{\text{TN+FP}}, \tag{13}$$

$$\text{FPR}=\frac{\text{FP}}{\text{FP+TN}}. \tag{14}$$

In addition, the maximum Recall achievable at a precision of at least 85% is calculated and reported. Per image evaluation and dataset level aggregation provide complementary perspectives on the performance of segmentation models [40],[41].

3.3. Implementation details

All experiments are conducted on a server equipped with four NVIDIA RTX 3090 GPUs (24 GB memory per GPU) to meet the computational and memory demands of model training. Model parameters are optimized using the Adam optimizer, which provides stable convergence and adaptive learning in deep learning tasks. The initial learning rate was set to 1×10^{-4} and kept constant throughout training. The batch size is set to 16 to balance resource utilization and gradient estimation stability. The model is trained for 150,000 iterations, with the entire training process completing in approximately 43 hours on the specified hardware. During diffusion inference, DPM-Solver is adopted for accelerated sampling. The diffusion timesteps are set to 1000, with 25 sampling steps using a first-order solver. The input for inspection is a single-channel infrared image of size $1\times 224\times 224$, consistent with the training setting.

4. Results and discussion

This section evaluates IR-SegDiff in terms of segmentation performance, sampling uncertainty, and computational efficiency. Quantitative and qualitative comparisons with established segmentation models assess defect localization and boundary delineation. Ensemble sampling and controlled noise experiments examine segmentation performance under simulated measurement noise and the utility of prediction disagreement for false-negative localization. Comparisons of sampling methods using identical model weights assess the trade-off between segmentation accuracy and inference cost. Ablation studies examine the contributions of conditional guidance and multiscale feature fusion to defect segmentation and background false-positive suppression.

4.1 Comparison of defect detection models

To systematically evaluate the performance of the proposed method in infrared defect segmentation, several widely used segmentation models are selected as baselines, including U-Net[42], U-Net++[43], Attention U-Net[44], DeepLabV3[45], DeepLabV3+[46], SegNet[47], and PSPNet[48]. These models represent typical segmentation architectures, encompassing encoder–decoder structures, attention-enhanced networks, and multi-scale contextual modeling approaches. All methods are evaluated using the same data splits and test inputs, with consistent preprocessing and evaluation procedures, to ensure a fair comparison of their capabilities in identifying defect regions in infrared images.

Table 2 presents the quantitative comparison of different methods on the constructed dataset. With a single sampling run, IR-SegDiff achieves an $F1$-Score of 69.72%, an IoU of 57.16%, and a Recall of 72.66%, exceeding all compared models. Averaging five independent sampling outputs further increases $F1$-Score, IoU, Precision, and Recall to 76.72%, 65.43%, 80.28%, and 75.04%, respectively. Among the baselines, U-Net achieves the highest $F1$-Score of 41.20%, whereas DeepLabV3 obtains the highest IoU of 27.86%; both remain below the corresponding IR-SegDiff results. Although U-Net achieves the highest Precision of 88.06%, its Recall is only 27.38%, indicating substantial omission of defect pixels. Overall, IR-SegDiff achieves the highest $F1$-Score, IoU, and Recall, while repeated sampling further improves all four segmentation metrics.

Table 2
Quantitative comparison with representative segmentation methods.

Method	*F*1-Score/%	IoU/%	Precision/%	Recall/%
U–Net	41.20	26.76	88.06	27.38
U–Net++	33.58	20.91	83.20	21.74
Attention U–Net	35.27	23.03	80.92	24.33
SegNet	35.48	22.19	62.89	25.74

PSPNet	28.71	17.25	80.92	17.89
DeepLabV3	41.05	27.86	82.63	30.07
DeepLabV3+	27.12	16.68	84.88	17.33
IR−SegDiff (Single run)	69.72	57.16	69.94	72.66
IR−SegDiff (ensemble of 5 runs)	**76.72**	**65.43**	**80.28**	**75.04**

As shown in Fig. 7, four representative infrared images exhibiting different defect morphologies are selected for a visual comparison of the segmentation results obtained by different methods. For circular defects, most methods can locate the main defect regions, although some predictions exhibit contour shrinkage or irregular boundaries. For defects with complex contours, such as star-shaped defects, several baseline methods exhibit local omissions and fail to preserve fine structural details. In samples containing elongated defects, fragmented predictions, shape deviations, and false positive detections in background regions can also be observed. In comparison, the predicted contours generated by IR-SegDiff show better agreement with the ground-truth masks, with relatively complete preservation of complex boundaries and elongated defect regions. These qualitative observations are consistent with the *F*1-Score and IoU comparisons reported in Table 2.

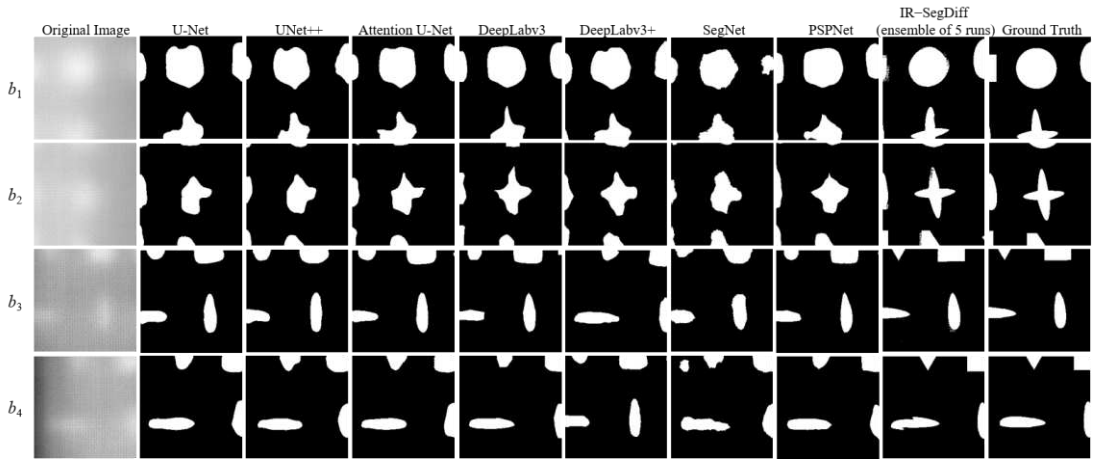


Fig. 7. Visual comparison of the segmentation performance of different models.

4.2 Noise robustness and uncertainty analysis

Measurement noise can obscure defect features in infrared images, leading to local segmentation errors and missed detections. This section examines the effect of ensemble sampling on segmentation performance and uses controlled noise experiments to assess defect identification at different noise levels. The utility of sampling uncertainty for false-negative localization is further evaluated using the false-negative capture rate at a fixed pixel review budget.

1) Effect of sampling ensemble size on segmentation performance

To evaluate the effect of averaging multiple sampling results on infrared defect segmentation, the number of sampling runs was set to K = 1, 10, and 20, while keeping the input images, model weights, and inference configuration unchanged. The continuous-valued segmentation outputs were averaged, and a threshold of 0.5 was applied to generate binary predictions. Table 3 presents the evaluation results across all test images. As the number of sampling runs increased from 1 to 10, the F1-Score improved from 69.72% to 76.67%, while IoU increased from 57.16% to 65.46%. Increasing the number of sampling runs to 20 further improved these metrics to 76.93% and 65.77%, respectively, although the additional gains were smaller. Meanwhile, Precision increased from 69.94% to 81.20%, and Recall improved from 72.66% to 75.03%. These results demonstrate that sampling ensembles improve segmentation performance on the current test set, with a more pronounced improvement in Precision, indicating their effectiveness in reducing false positive detections. However, the performance gains diminish as the number of sampling runs increases.

Table 3

Effect of ensemble sampling count on infrared defect segmentation performance.

Number of Samples, K	F1-Score/%	IoU/%	Precision/%	Recall/%
1	69.72	57.16	69.94	72.66
10	76.67	65.46	80.58	74.87
20	76.93	65.77	81.20	75.03

Fig. 8 presents the segmentation results for four representative input images. For input images b_2 and b_4, single sampling produces numerous scattered false positive predictions, making it difficult to clearly delineate the complete defect contours. After averaging multiple sampling results, these scattered false positives in the background are substantially reduced, and the predicted regions exhibit better agreement with the ground-truth defect morphologies. At K = 20, the F1-Scores for these two input images increase from 32.82% and 35.99% to 81.22% and 92.04%, respectively. In contrast, the single-sampling predictions for cases b_1 and b_3 already show relatively good agreement with the ground truth, and increasing the number of sampling runs yields only limited improvements. These observations indicate that the benefits of sampling ensembles are input-dependent.

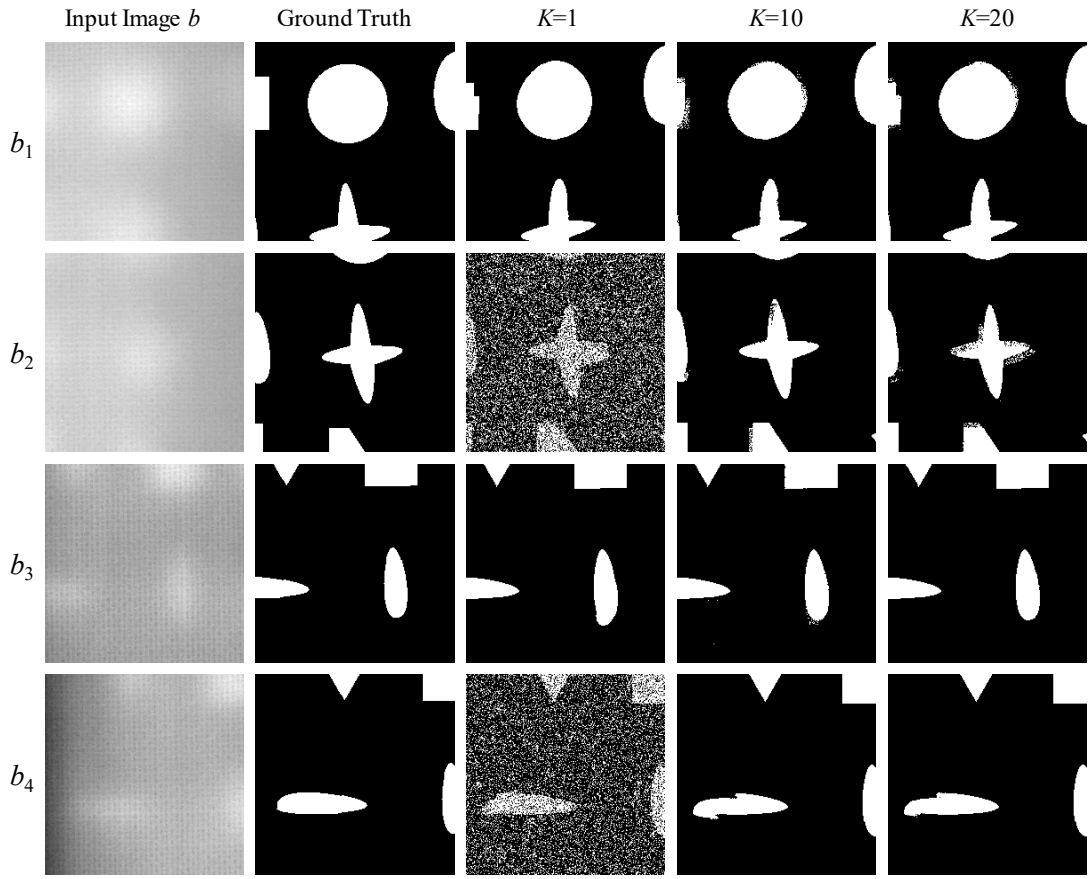


Fig. 8. Qualitative comparison of defect segmentation with different sampling.

2) Effect of simulated measurement noise on defect segmentation

To investigate the effects of measurement disturbances on defect identification, fixed local regions of specimens C4 and C5 are selected, and zero-mean Gaussian noise with standard deviations of σ =5,10,15 grayscale levels is added. Images without additional noise serve as the baseline. The input regions, defect annotations, and model parameters are kept unchanged, and three independent noise realizations are generated at each noise level. Table 4 compares the segmentation performance of single sampling (K=1) and 20-sample (K=20) ensembles, with the mean and sample standard deviation across the three realizations reported for each noise level. Fig. 9 presents the input images, segmentation results, sampling standard deviation maps, and error distributions for the same noise realization index, enabling a visual comparison of defect identification performance under different levels of interference.

Table 4

Segmentation performance under controlled gaussian noise.

Case	Noise SD, σ	$F1$-Score /%, K=1	$F1$-Score /%, K=20	Recall /%, K=1	Recall /%, K=20
C4	0	91.52	91.38	90.60	90.46

	5	91.27 ± 0.67	91.63 ± 0.48	89.93 ± 0.64	90.92 ± 0.81
	10	87.88 ± 2.16	89.03 ± 3.01	84.90 ± 3.00	86.20 ± 3.89
	15	70.72 ± 8.60	79.10 ± 6.85	59.85 ± 10.23	70.38 ± 9.21
C5	0	54.29	89.79	66.25	90.32
	5	54.39 ± 0.02	89.80 ± 0.17	66.39 ± 0.03	91.27 ± 0.80
	10	54.74 ± 0.08	89.46 ± 0.15	66.65 ± 0.17	91.75 ± 0.40
	15	53.75 ± 0.15	87.21 ± 1.22	64.90 ± 0.16	87.94 ± 1.50

Table 4 shows that ensemble segmentation performance under low-intensity noise remains close to that obtained from the original inputs, whereas stronger disturbances affect the two cases to different extents. For C4, as the noise standard deviation σ increases to 15, the $F1$-Score of the ensemble prediction decreases from 91.38% to 79.10%, while Recall decreases from 90.46% to 70.38%. As illustrated in Fig. 9, the predicted defect region exhibits shrinkage and local omissions under high-noise conditions, accompanied by an expansion of the false negative regions in the error maps. In contrast, the ensemble prediction for specimen C5 remains relatively stable under the applied noise, maintaining an $F1$-Score of 87.21% and a Recall of 87.94% even at the highest noise level of σ=15. The corresponding visual errors are mainly characterized by boundary deviations and local morphological distortions. These differences suggest that the effects of measurement disturbances depend not only on noise intensity but also on the distinguishability of defect features in the input images.

Based on these observations, sampling ensembles improve the recovery of defect regions under noisy conditions. For specimen C4 at σ=15, increasing the number of sampling runs to 20 improves the $F1$-Score from 70.72% to 79.10% and Recall from 59.85% to 70.38%, consistent with the more complete preservation of defect regions observed in Fig. 9. For specimen C5, ensemble predictions effectively reduce scattered false positives in single-sampling outputs across all noise levels, demonstrating their ability to suppress sampling variability. Meanwhile, the standard deviation maps reveal pronounced prediction disagreement along defect boundaries and within certain noise-affected regions, providing spatial information about prediction stability. These results indicate that repeated diffusion sampling can provide both ensemble segmentation predictions and pixel-wise variability estimates, offering complementary information for defect identification under measurement disturbances and the subsequent localization of regions requiring further inspection.

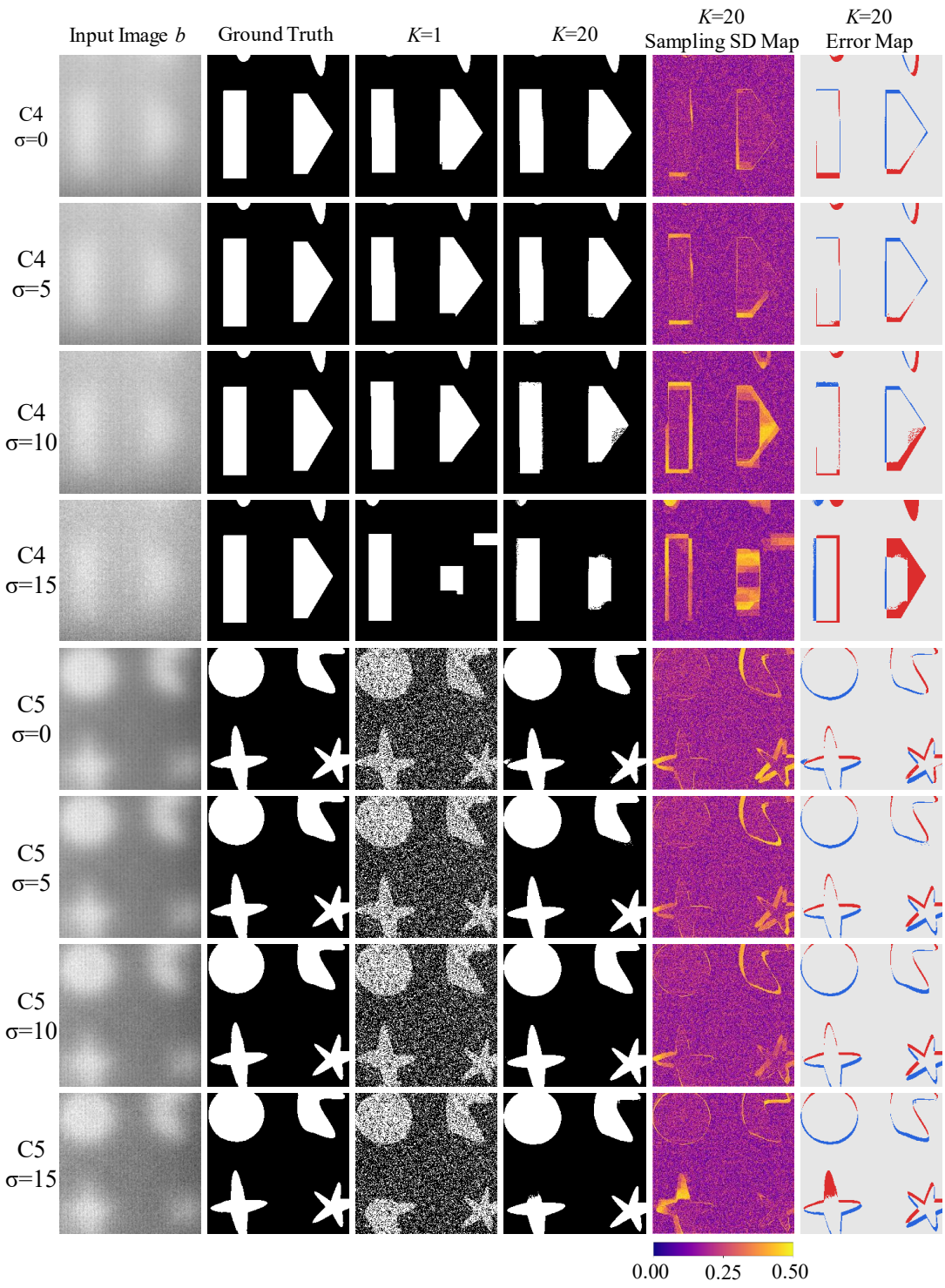


Fig. 9. Defect segmentation and sampling uncertainty under simulated Gaussian noise. In the error maps, red, blue, and light gray indicate false negatives, false positives, and correctly classified pixels, respectively.

3) Sampling uncertainty for False-Negative Localization

To quantify the utility of sampling uncertainty for false-negative (FN) localization, pixels were ranked by their sampling standard deviation, and the top 10% were selected as priority review regions. The threshold margin baseline selected an equal number of

pixels whose mean prediction scores were closest to 0.5. The false-negative capture rate is defined as the fraction of all false-negative pixels covered by the selected regions, measuring localization performance at a fixed pixel review budget. Each input image was evaluated separately. Table 5 summarizes the results evaluated independently for each input image, with the mean and standard deviation across three noise realizations reported for the noisy conditions.

Table 5

False-Negative capture at a 10% pixel review budget.

Case	Noise SD, σ	FN Capture /%, Sampling SD	FN Capture /%, Threshold Margin
C4	0	33.24	31.47
	5	29.53 ± 6.95	27.27 ± 7.16
	10	54.01 ± 6.35	52.92 ± 7.38
	15	34.09 ± 6.61	33.33 ± 6.64
C5	0	59.49	57.72
	5	55.45 ± 5.74	54.24 ± 6.14
	10	51.59 ± 9.28	50.29 ± 9.99
	15	50.14 ± 9.59	48.61 ± 9.72

As shown in Table 5, sampling standard deviation achieves a higher mean false negative capture rate than the threshold-distance baseline across all noise levels for both test specimens. For C4 at σ=5, the capture rate was 29.53%, compared with 27.27% for the baseline. For C5 at the highest noise level, the corresponding rates were 50.14% and 48.61%, indicating that the selected 10% of image pixels covered approximately half of all false-negative pixels. These results suggest that repeated-sampling disagreement provides complementary localization information beyond mean prediction scores. Together with the preceding segmentation results, they show that diffusion sampling supports both improved defect segmentation through ensemble averaging and the prioritization of regions for review through pixelwise prediction disagreement. Consequently, the model output extends beyond binary defect localization to include information on prediction stability, providing a basis for defect interpretation and targeted manual review under measurement disturbances.

4.3 Comparison of sampling methods

To evaluate the effects of different sampling strategies on segmentation accuracy and computational efficiency, this study compares 1000-step DDPM, 100-step second-order DPM-Solver, and 25-step first-order DPM-Solver under identical model weights and test settings. All methods employ an ensemble of five independent sampling runs. DDPM generates segmentation results through sequential reverse denoising, whereas

DPM-Solver exploits the continuous-time formulation of the diffusion process to reduce the number of model evaluations required for reverse sampling through numerical integration.

As shown in Table 6, DDPM achieves slightly higher segmentation accuracy, but its sequential denoising process incurs substantial computational overhead. The 25-step first-order DPM-Solver achieves a 35.57-fold speed up. *F*1-Score decreases slightly from 76.98% to 76.72%, and IoU from 65.90% to 65.43%. These results suggest that the conditional spatial information supports defect reconstruction with fewer sampling updates, while ensemble averaging helps reduce variability across individual predictions. The 100-step second-order configuration provides no improvement over the 25-step first-order configuration in this evaluation, suggesting that additional sampling steps and a higher solver order do not necessarily improve segmentation performance. The 25-step first-order DPM-Solver is therefore adopted as the default configuration, offering a favorable balance between segmentation accuracy and computational efficiency.

Table 6

Comparison of performance across different sampling methods.

Ensemble of 5 Runs	F1-Score/%	IoU/%	Precision/%	Recall/%	Time/s	Speed-up
DDPM	76.98	65.90	80.96	75.16	260.05	1.00×
DPM-Solver (100step-2order)	76.62	65.35	80.25	74.89	27.59	9.43×
Ours	76.72	65.43	80.28	75.04	7.31	35.57×

4.4 Ablation study

To evaluate the effectiveness of the key modules incorporated into IR-SegDiff, four model variants are constructed, as summarized in Table 7. IR-SegDiff denotes the complete model incorporating the conditional encoder, CCAFM and MSFFM. IR-SegDiff-1 is obtained by removing MSFFM from the complete model, whereas IR-SegDiff-2 excludes CCAFM. IR-SegDiff-3 removes all three modules and serves as the baseline model. All variants employ identical data splits, image preprocessing procedures, and testing protocols. Furthermore, the same test images, random cropping locations, normalization parameters, five-sample diffusion ensembles, and 25-step DPM sampling are used to ensure a fair comparison.

Table 7

Ablation studies of models composed of different modules.

Model	Conditional encoder	CCAFM	MSFFM
IR-SegDiff	√	√	√

IR-SegDiff-1	√	√	×
IR-SegDiff-2	√	×	√
IR-SegDiff-3	×	×	×

As shown in Table 8, the complete IR-SegDiff model achieves the highest Precision and AP, reaching 80.28% and 77.65%, respectively. Removing MSFFM reduces AP from 77.65% to 74.82%, indicating that multi-scale feature fusion contributes to improved ranking of defect probabilities and pixel-wise discrimination. When CCAFM is removed, the *F*1-Score, IoU, and Recall increase, whereas precision and AP decrease. This suggests that CCAFM primarily suppresses low-confidence background responses and promotes more conservative predictions, rather than consistently improving the spatial overlap between predicted and ground-truth defect regions. Compared with the baseline model, the complete model achieves AP increases from 73.66% to 77.65%, indicating that the combination of conditional representations and the proposed feature fusion modules enhances the overall quality of probabilistic predictions.

Table 8

Standard segmentation performance of different ablation variants.

Model	*F*1-Score/%	IoU/%	Precision/%	Recall/%	AP/%
IR-SegDiff	76.72	65.43	**80.28**	75.04	**77.65**
IR-SegDiff-1	76.18	63.94	79.66	74.58	74.82
IR-SegDiff-2	**78.33**	**66.37**	79.27	**78.42**	76.78
IR-SegDiff-3	77.30	64.87	77.41	78.00	73.66

The advantages of the complete model become more evident under precision-oriented evaluation. As shown in Table 9, IR-SegDiff achieves the highest micro-averaged precision, $F_{0.5}$-Score, and specificity, reaching 83.51%, 81.64%, and 97.03%, respectively, while attaining the lowest FPR of 2.97%. Compared with the baseline model, the complete model improves micro-averaged precision and the $F_{0.5}$-Score by 5.05% and 3.26 %, respectively, while reducing FPR from 4.31% to 2.97%. At a precision threshold of at least 85%, IR-SegDiff maintains a Recall of 73.21%, outperforming the two ablated variants, whereas the baseline model fails to meet this precision requirement. These results indicate that the conditional encoder provides image-specific thermal response information to guide the diffusion-based segmentation process, MSFFM integrates defect features across multiple scales, and CCAFM helps suppress spurious responses in background regions. The combined contributions of these three modules primarily enhance the reliability of positive predictions and the suppression of false positives, making IR-SegDiff particularly suitable for infrared defect detection applications where a low false positive rate is essential.

Table 9

Precision-oriented performance and false-positive suppression.

Model	Micro Precision/%	$F_{0.5}$-Score /%	Specificity/%	FPR/%	Recall at Precision ≥ 85%/%
IR-SegDiff	**83.51**	**81.64**	**97.03**	**2.97**	**73.21**
IR-SegDiff-1	81.96	80.33	96.71	3.29	69.48
IR-SegDiff-2	80.94	80.38	96.30	3.70	64.80
IR-SegDiff-3	78.46	78.38	95.69	4.31	—

Note: "—" indicates that the model did not reach 85% precision on the test-set Precision–Recall curve.

5. Conclusions

This study proposes IR-SegDiff, a conditional diffusion-based segmentation model for subsurface defect segmentation and assisted interpretation in aerospace structural materials using infrared thermography. By incorporating a conditional encoder, conditional CCAFM, and MSFFM, the proposed model integrates multi-scale information from infrared images into the diffusion denoising process, enabling joint representation of defect structures and contextual features. On a test set partitioned by physical specimens, the five-sample ensemble achieves an $F1$-Score of 76.72% and an IoU of 65.43%, outperforming seven representative segmentation networks. Ablation experiments further demonstrate the advantages of the complete model in terms of precision, average precision, and suppression of false positive detections in background regions. By employing a 25-step DPM-Solver, the sampling time per image for the five-sample ensemble is reduced from 260.05 s with DDPM to 7.31 s, achieving a 35.57-fold speedup while maintaining comparable segmentation performance. Controlled noise experiments demonstrate that averaging repeated sampling results improves defect identification under measurement disturbances, while pixel-wise sampling standard deviation provides additional information for localizing missed defects beyond the mean predictions. At the highest noise level, the 10% of pixels selected according to sampling standard deviation capture 34.09% and 50.14% of false negative pixels in the two test cases, respectively, highlighting the value of prediction disagreement for prioritizing regions requiring further inspection. By integrating condition-guided segmentation, efficient diffusion inference, and uncertainty characterization, the proposed method extends infrared defect detection beyond a single segmentation mask to provide additional information on prediction stability. This offers a unified framework for automated subsurface defect identification and targeted manual review in aerospace materials.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. U23A20111), the Ningbo Yongjiang Talent Introduction Programme (No. 2021A-042-G) and the Ningbo International Sci-tech Cooperation Projects (No. 2024H009).

References

[1] Zou Q, Huang R, Mu X, et al. Breaking Through Flutter Barrier of Rigid-Elastic Coupling Aircraft[J]. Applied Mechanics Reviews, 2026, 78(2): 020801.

[2] Ogunleye R O, Rusnáková S, Javořík J, et al. Advanced sensors and sensing systems for structural health monitoring in aerospace composites[J]. Advanced Engineering Materials, 2024, 26(22): 2401745.

[3] Xi T, Wang P, Shi Y, et al. Characterization and Quantitative Evaluation of Surface Cracks on Steel Plates Based on Eddy Current Pulsed Thermography[J]. IEEE Transactions on Instrumentation and Measurement, 2025.

[4] Tulshibagwale N, Brodnik N, Muir C, et al. Acoustic emission in ceramic matrix composites[J]. Applied Mechanics Reviews, 2024, 76(4): 040801.

[5] Wang G, Chen Z, Yan J, et al. Peculiar Functions and Potential Applications of Acoustic Metamaterials in Wave Manipulation: A Review[J]. Applied Mechanics Reviews, 2025: 1-108.

[6] Cui W, Song K, Zhang Y, et al. Fine-grained tiny defect detection in spiral welds: a joint framework combining semantic discrimination and contrast transformation[J]. IEEE Transactions on Instrumentation and Measurement, 2025.

[7] Gao Q, Wu M, Qin X, et al. Machine vision driven magnetic particle inspection technology: principles, applications and trends[J]. Measurement Science and Technology, 2026, 37(3): 032001.

[8] Wang H, Chen H, He Y, et al. Research on CFRP small-gap defect detection by joint scanning laser thermography based on restored pseudo heat flux[J]. IEEE Transactions on Instrumentation and Measurement, 2025.

[9] Li R, Wang F, Yang F, et al. Investigation on the detection of crack defects in metals using chirp ultrasonic induced infrared thermography[J]. IEEE Transactions on Instrumentation and Measurement, 2025.

[10] Hsiao T Y, Sfarra S, Liu Y, et al. Two-dimensional Hilbert-Huang transform-based thermographic data processing for non-destructive material defect detection[J]. Quantitative InfraRed Thermography Journal, 2025, 22(4): 297-312.

[11] Wang F, Wang Y, Liu J, et al. The feature recognition of CFRP subsurface defects using low-energy chirp-pulsed radar thermography[J]. IEEE Transactions on Industrial Informatics, 2019, 16(8): 5160-5168.

[12] Das P, Arora V, Mulaveesala R. Wavelet transform based post processing approach for pulse compression favourable frequency modulated thermal wave imaging for non-destructive testing and evaluation[J]. Infrared Physics & Technology, 2025, 148: 105843.

[13] Lu X, Yi Q, Tian G Y. A comparison of feature extraction techniques for delamination of CFRP using eddy current pulse-compression thermography[J]. IEEE Sensors Journal, 2020, 20(20): 12415-12422.

[14] Roche J M, Leroy F H, Balageas D. Images of TSR coefficients: A simple way for a rapid and efficient detection of defects[J]. Materials evaluation, 2014, 72(1): 73-82.

[15] Wang F, Liu J, Dong B, et al. Optimization of thermal-wave radar thermography by transverse heat flow suppression technique for accurate defect detection of

CFRP laminates[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-10.
[16] Liu Y, Yao Y, Wang F, et al. Review of unsupervised machine learning methods in active infrared thermography for defect detection and analysis[J]. Quantitative InfraRed Thermography Journal, 2025: 1-28.
[17] Yao N, Guo Y, Zhao Y, et al. Subsurface Defect Detection in Carbon Fiber-Reinforced Plastics Using Infrared Polarization Imaging[J]. IEEE Transactions on Instrumentation and Measurement, 2025, 74: 1-16.
[18] Cheng X, Chen P, Wu Z, et al. Automatic detection of CFRP subsurface defects via thermal signals in long pulse and lock-in thermography[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-10.
[19] Matarrese T, Marani R, Palumbo D, et al. Influence of lock-in thermography set-up parameters on the capability of a temporal convolutional neural network to characterize defects in a CFRP[J]. Optics and Lasers in Engineering, 2024, 182: 108455.
[20] Morelli D, Marani R, D'Accardi E, et al. A convolution residual network for heating-invariant defect segmentation in composite materials inspected by lock-in thermography[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-14.
[21] Lim W, Sfarra S, Hsiao T, et al. Physics-Informed Neural Networks for Defect Detection and Thermal Diffusivity Evaluation in Carbon Fiber-Reinforced Polymer Using Pulsed Thermography[J]. IEEE Transactions on Instrumentation and Measurement, 2025, 74, 1-10.
[22] Li X, Ying X, Zhu W, et al. Nondestructive detection and analysis based on data enhanced thermography[J]. Measurement Science and Technology, 2022, 33(6): 064006.
[23] Fang Q, Ibarra-Castanedo C, Yuxia D, et al. Defect enhancement and image noise reduction analysis using partial least square-generative adversarial networks (PLS-GANs) in thermographic nondestructive evaluation[J]. Journal of Nondestructive Evaluation, 2021, 40(4): 92.
[24] Sabato A, Dabetwar S, Kulkarni N N, et al. Noncontact sensing techniques for AI-aided structural health monitoring: a systematic review[J]. IEEE Sensors Journal, 2023, 23(5): 4672-4684.
[25] Cheng L, Kersemans M. Dual-IRT-GAN: A defect-aware deep adversarial network to perform super-resolution tasks in infrared thermographic inspection[J]. Composites Part B: Engineering, 2022, 247: 110309.
[26] Cheng L, Tong Z, Xie S, et al. IRT-GAN: A generative adversarial network with a multi-headed fusion strategy for automated defect detection in composites using infrared thermography[J]. Composite Structures, 2022, 290: 115543.
[27] Zhou X, Zhang Y, Ren Z, et al. DiffDD: A surface defect detection framework with diffusion probabilistic model[J]. Advanced Engineering Informatics, 2024, 62: 102637.
[28] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[29] Croitoru F A, Hondru V, Ionescu R T, et al. Diffusion models in vision: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(9): 10850-10869.
[30] Xia Z, Zhao Y, Gu J, et al. The end-to-end chip surface defect segmentation method based on the diffusion model and attention mechanism[J]. Engineering Applications of Artificial Intelligence, 2025, 155: 111131.
[31] Fang T, Zhang M, Song R, et al. Human-Guided Data Augmentation via Diffusion Model for Surface Defect Recognition Under Limited Data[J]. IEEE Transactions on Instrumentation and Measurement, 2025.
[32] Wei Y, Zhang Y, Xiao Y, et al. IRT-diffusion: combined denoising diffusion probabilistic models with thermal signal processing methods for automated defect detection in composites using infrared thermography[C]//International Conference on Optical and Photonic Engineering (icOPEN 2024). SPIE, 2025, 13509: 435-439.
[33] Luo Z, Wang H, Huang Y, et al. Nondestructive detection of CFRP subsurface defects using transient lock-in thermography[J]. Journal of Thermal Analysis and Calorimetry, 2025, 150(11): 8229-8239.
[34] Liu G, Liu W, Gao W, et al. Nonlinear frequency modulation thermography for defect detection in CFRP composites based on deep learning YOLO-V8 network model[J]. Nondestructive Testing and Evaluation, 2025: 1-13.
[35] Gao T, Hu W, Chen M, et al. MSDMAT-BTS: Multi-scale diffusion model and attention mechanism for brain tumor segmentation[J]. Biomedical Signal Processing and Control, 2025, 104: 107505.
[36] Lu C, Zhou Y, Bao F, et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps[J]. Advances in neural information processing systems, 2022, 35: 5775-5787.
[37] Chen Z, Lu S. Caf-yolo: A robust framework for multi-scale lesion detection in biomedical imagery[C]//ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025: 1-5.
[38] Taha A A, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool[J]. BMC medical imaging, 2015, 15(1): 29.
[39] Margolin R, Zelnik-Manor L, Tal A. How to evaluate foreground maps?[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 248-255.
[40] Csurka G, Larlus D, Perronnin F, et al. What is a good evaluation measure for semantic segmentation?[C]//Bmvc. 2013, 27(2013): 10-5244.
[41] Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation[J]. BMC research notes, 2022, 15(1): 210.
[42] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015: 234-241.
[43] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//International workshop on deep learning in medical image analysis. Cham: Springer International Publishing, 2018:

3-11.
[44] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.
[45] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.
[46] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.
[47] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
[48] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.